DOCUMENT RESUME

ED 150 176 TM 006 891

AUTHOR Northrop, Lois C.
TITLE Experimental Item Types to Measure Judgment.
INSTITUTION Civil Service Commission, Washington, D.C. Personnel Measurement Research and Development Center.
REPORT NO PRDC-TM-77-14
PUB DATE Aug 77
NOTE 13p.; For related document, see TM 006 890

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Abstract Reasoning; Administrative Personnel; Comprehension; *Decision Making Skills; Item Analysis; *Measurement Techniques; Occupational Tests; Problem Solving; Professional Personnel; *Psychological Tests; *Test Construction; Testing Problems; *Test Items; Test Reliability; Test Validity; Verbal Ability
IDENTIFIERS *Judgment; *Professional Administrative Career Exam

ABSTRACT

The Professional and Administrative Career Examination measures judgment ability using comprehension and logical order of events item types. Since the verbal component of comprehension items is high, and since these items are extremely difficult to write and document, a search was conducted for a simpler item type for assessing judgment. Three alternative item types were investigated: word selection, evaluation of comparisons, and analogies. It was concluded that no clearly satisfactory replacement for the comprehension item types has been found. Item statistics for word selection and evaluation of comparisons were very similar to those for comprehension; while analogy items were more strongly associated with verbal and reasoning abilities than with the other item types which assess judgment ability. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150176

Personnel
Research and
Development
Center

Technical Memorandum 77-14

Experimental Item Types to Measure Judgment

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 891

United States
Civil Service Commission
Bureau of Policies and Standards

Technical Memorandum 77-14

EXPERIMENTAL ITEM TYPES TO MEASURE JUDGMENT

Lois C. Northrop

Personnel Research and Development Center
United States Civil Service Commission
Washington, D. C. 20415
August 1977

EXPERIMENTAL ITEM TYPES TO MEASURE JUDGMENT

ABSTRACT

This report summarizes experiences with the Comprehension item type used in Test 500 (PACE) to assess the Judgment ability construct and discusses a resultant search for a new item type with which to replace it.

Three new item types were proposed and tried out experimentally in three series of Test 500. Item analysis results of this try out are reported and compared with data for the Comprehension items used operationally in the same three series. A discussion of these results and some conclusions are also presented.

CONTENTS

	Page
Comprehension Item Type in PACE	1
The Search for a New Item Type	1
The Three Alternative Item Types Selected	2
Word Extensions (Selection)	2
Evaluation of Comparisons	3
Analogies	4
Experimental Tryout	4
Results	5
Conclusions	5
References	8

EXPERIMENTAL ITEM TYPES TO MEASURE JUDGMENT

Judgment is an ability measured in the written test for the Professional and Administrative Career Examination (PACE) using the Comprehension and Logical Order of Events item types. Experience with writing the Comprehension item type and data concerning its correlation with the test part used to measure verbal ability prompted a search for alternate item types to measure judgment. This search led to the selection of three experimental item types, any one of which could replace the Comprehension item type. The literature review and the rationale for these new item types are discussed in some detail. An extensive literature review pertaining to the existence of the judgment ability is reported in a separate paper (Northrop, 1976).

Comprehension Item Type in PACE

In spite of the many stipulations, precautions and measures to increase objectivity, and to reduce vagueness and verbal loadings used by item writers, PACE Comprehension items are a source of concern for two reasons. First, the Verbal component is high. The correlation between Verbal ability items and Comprehension items is approximately .67.[1] (The correlations between the Comprehension items and the other ability measures in the written test are less--Comprehension and Induction is about .44, Comprehension and Deduction is about .59, Comprehension and Numerical is about .45. With the exception of Verbal, and possibly Deduction, there should be little concern about the size of these correlations since it is expected that most abilities are moderately correlated. As Adkins and Lyerly (1952) have pointed out, factorially pure tests of any kind are virtually impossible to achieve since the medium of the test--words, figures, and numbers--intrudes. Second, the items have been extremely difficult to write and to document. As a result, the great majority of the items have been rejected either in the content review stage because of concern over possible ambiguities or in the test assembly stage because of item statistics.

The Search for a New Item Type

A search has been conducted for a "new" and simpler item type with which to assess judgment ability. Specifically, an item type which would reduce the verbal comprehension variance was sought. At the same time it had to satisfy criteria for making judgments: a problem must be presented wherein all of the necessary facts are not given; the given data must be analyzed and then integrated with relevant information from a pool of general knowledge; a conclusion and then a decision must be made concerning the best solution.[2]

A review of the research done by Adkins and Lyerly (1952) and by Guilford and his numerous associates (Guilford, Hertzka, Berger and Christensen, 1952; Guilford, Hertzka and Christensen, 1951; Nihira, Guilford, Hoepfner and Merrifield, 1964) suggested several possible appropriate measures of judgment. In a discussion of the separation of evaluative (judgment) and cognitive abilities Nihira et al. (1964, p. 6)

---

[1] As discussed in The definition and measurement of judgment (Northrop 1976), the Judgment factor is complex. Acts of judgment pervade most problem-solving tasks and measures of diverse abilities (verbal, reasoning, creativity, etc.) may have loadings on a Judgment factor.

[2] Operationally, in the PACE, judgment has been defined as "...the ability to solve a presented problem when all the necessary facts to solve the problem are not given. Solution to the problem involves making some reasonable assumptions or anticipating what the most likely of several possible occurrences might be..." (U. S. Civil Service Commission, 1974)

state that "a test designed to measure an evaluative ability should pitch the level of the evaluative task so that individual differences in test scores reflect that ability." Verbal aspects of the test should be made so easy that they contribute little or nothing to the variance. Recall of specific and/or detailed factual information should not be a requirement. The difficulty in the evaluative task is introduced by providing an appreciable degree of uncertainty as to which alternative is correct. All alternatives can be made relevant, desirable and/or acceptable under standards of poor judgment. Verbal and recall tasks can be minimized by stating the problem clearly and simply and using general and familiar information.

In searching for alternative item types, one question which needed to be resolved was "does judgment transcend the medium of the test?" In other words, are the mental processes involved in judgment the same whether the test medium is symbolic, numerical or verbal?

After examining and comparing many tests of all three varieties (symbolic, numerical and verbal) relative to their factor loadings, their intercorrelations with other measures in the same battery, and their reliabilities, it became evident that verbal item types were the only ones suitable if the PACE definition of judgment was to be satisfied and if sufficient variety of subject matter was to be available. Figural and symbolic item types were eliminated quickly as being too restrictive (too specific) in what they measured.

By processes of comparison, evaluation and elimination, three item types were finally selected for trial. Statistically they were considered the most suitable for the ability they were to measure and additionally it was hoped that within these three formats, items which were verbally simple but conceptually difficult could be constructed.

The Three Alternative Item Types Selected

Word Selection

The first item type chosen was called Word Extensions (Nihira et al, 1964) or Word Selection (Guilford et al, 1953; Adkins and Lyerly 1952, Guilford and Hoepfner, 1971).

Example: A box always has:

A) wood
B) size
C) a lid
D) nails
E) a label

The task is to select the alternative which denotes what the item always possesses or always involves. Every alternative could be correct but only one is always correct; an act of evaluation is a necessity. In the above example B is correct - a box always has a size; it may or may not be of wood, or have nails, a lid or a label. In the Adkins and Lyerly (1952) study Word Selection loaded very minimally on two Reasoning factors -- Hypothesis Verification (.27) (possibly allied with the Air Force J factor) and Concept Formation (.18). Adkins and Lyerly proposed that the Hypothesis Verification factor represented more than merely induction. Respondents might have been approaching the solution to the problem by evaluating all hypotheses suggested to them by the alternatives - i.e., they were verifying, rather than simply educing, a rule or principle. A test called Practical Situations (Practical Judgment) loaded .39 on Hypothesis Verification, and also correlated .39 with Word Selection. Additionally, a correlation of .49 between Word Selection and Sentence Order (Logical Order of Events) was reported, although the latter test was more strongly associated (.50) with a Verbal Relations factor than with reasoning. Verbal loadings for Word Selection were not statistically significant. The purpose of the Adkins and Lyerly investigation was a factor analysis of Reasoning tests. No specific Judgment factor emerged in their study.

Word Selection was included in a battery of Evaluative Ability tests by Guilford et al. in 1953. Its alternate form reliability was .61, considerably better than reported for such reference tests as Practical Judgment (.43). Correlations of .40 to .50 were found between Word Selection and Logical Classification (of statements into one of five categories), Vocabulary, and Logical Reasoning (syllogisms). Highest loadings for Word Selection were on the factors Logical Evaluation (equivalent to Logical Reasoning) (.47) and Verbal Relations (.36), the two most general factors to emerge. Again a Judgment factor failed to appear; the test intended as its marker was associated most strongly with General Reasoning (loading = .42) and correlated most highly with Syllogisms (.30).

An analysis of Semantic Evaluative abilities by Nihira et al (1964) again showed an association between Word Extension and a Verbal Relations factor (Evaluation of Semantic Relations; a loading of .41). The Evaluation of Semantic Implications factor contained both Word Extensions (loading = .37) and Commonsense Judgment I (select the two best reasons why a plan is faulty; loading = .36). Reliabilities for these two tests were consistent with previous findings (.64 and .40 respectively). Their correlation with each other was .34.

Word Selection (Extensions) did exhibit some cognitive variance (loadings of .25 and .38 on one of six cognitive factors in two different samples of high school students) and showed moderate correlations (.40 to .50) with several reading and vocabulary measures, but this was predictable since the test medium is verbal. The item type was chosen as a possibility for assessing judgment because of the consistent data available from several different studies and because it seemed possible to construct items such that all of the mental processes involved in judgment would be utilized in selecting the correct alternative. It was thought that reading comprehension variance would be minimized, since both stem and alternatives would consist of one or two words; vocabulary would not be an issue if complexity of concept rather than word meaning was emphasized. The item type seemed to lend itself well to the criterion of solving a problem when all of the facts are not given, thereby necessitating an application of one's general knowledge for a solution.

Evaluation of Comparisons

A second item type considered to be appropriate for the assessment of judgment comes from a measure called Evaluation of Comparisons (Guilford et al., 1953; Guilford and Hoepfner, 1971).

Example: (two words representing objects or concepts to be compared)

X - candle
Y - light bulb

Which of the following statements is the most justifiable comparison?

A) X was invented earlier than Y.
B) X is more useful than Y.
C) X is easier to obtain than Y.
D) X is more breakable than Y.
E) X gives off more light than Y.

Comparisons B, C, D, and E are not satisfactory - each may or may not be true, depending on specific situations or circumstances. Comparison A is true, and is therefore the most justifiable comparison to be made. It is obvious that some additional experience or knowledge is required for making a decision about the correct answer.

This test was considered to be rather complex and therefore not a particularly good measure of any of Guilford's specific evaluative abilities. It did have a loading of .36 on the more general factor Logical Evaluation as well as a small loading (.20) on the Verbal factor. As did Word Selection, Evaluation of Comparisons correlated with several other verbal tests including Vocabulary (.45) and Reading (.37). Similar correlations were obtained between Evaluation of Comparisons and a number of tests of a syllogistic nature. Furthermore, Word Selection and Evaluation of Comparisons correlated .45 with each other.

The reliability of Evaluation of Comparisons was .56, higher than that for traditional Practical Judgment marker tests.

While the statistical properties of this item type did not appear to be particularly robust, it was reasonable to infer that it would be useful for assessing judgment in the written test for PACE. An analysis of two concepts is required (as opposed to only one in Word Selection) and a relationship between the two must be established. Subsequently, five explicit judgments must be analyzed and the correct conclusion derived through the application of outside knowledge of a general academic or experiential nature. Items can be written with a wide range of difficulty depending upon the complexity of concepts chosen for comparison. It was thought that reading comprehension and verbal (vocabulary) variance could be minimized as for Word Selection.

Analogies

The third item type considered for measuring judgment is a form of analogy in which the relationship expressed by the first pair of words is fairly obvious but the choice of a correct alternative to demonstrate the same relationship in the second pair is more difficult.

Example: EVIDENCE: CRIMINAL as WEB:

A) thread
B) spider
C) fly
D) net
E) trap

The correct answer is C. As EVIDENCE is used to capture a CRIMINAL so a WEB is used by the spider to capture a FLY. The other four alternatives are also related to WEB but not in the same way.

An analogies test of this variety was included in the battery used by Adkins and Lyerly (1952) in their study of reasoning abilities. It had a loading of .32 on the factor Hypothesis Verification (it will be recalled that Word Selection also appeared in this factor). The highest loading for Analogies was .33 on Concept Formation, which is probably associated with an ability to generalize or to arrive at abstract generalizations. A loading of .26 was obtained for the factor Deduction and .21 for Perception of Abstract Similarities. A correlation of .41 was reported between Analogies and Practical Judgment; .65 between Analogies and Word Selection.

Nihira et al. (1964) included Analogies in their test battery for Semantic Evaluation Abilities. It appeared significantly (a loading of .53) in the Evaluation of Relations factor and had small loadings (.21 and .19) on the Evaluation of Classes and Evaluation of Implications factors. Loadings on two of the cognitive factors were .20 and .25. Its reliability was .60. Correlations with two verbal tests (Synonyms and Vocabulary) were .28 and .36; correlations with Logical Reasoning (Syllogistic form), Commonsense Judgment (two reasons why a plan is faulty) and Word Extensions (Word Selection) were .47, .32 and .51 respectively.

The Analogy item type seemed to have good potential for assessing judgment. Unlike Evaluation of Comparisons, respondents are not presented with five qualitative or quantitative statements to evaluate, but are required to figure out for themselves the relationships of the alternatives to the stem and decide which of the five choices is most appropriate. Frequently an awareness of the many types of existing relationships will help a competitor solve an analogy problem. This would not be measuring judgment, but this effect could hopefully be minimized by making the relationship in the first pair of words a fairly obvious one. As for the first two item types discussed, an extensive vocabulary is not a prerequisite for the successful completion of the analogy. Extrinsic general knowledge of the concepts presented in the word pairs may be required, in order that the precise relationship presented in the first half of the analogy is duplicated in the second half.

Reading comprehension variance for this item type could, of course, be held to a minimum. A wide range of item difficulty should also be possible through careful choice of concepts to be compared.

Experimental Tryout

These three item types were tried out experimentally with the written test used in PACE. Many items of each type were written by the staff of the Test Services Section in accordance with the subject-matter categories currently used for the Comprehension item type. The author picked the best of these items and edited them as required. The PACE test review team composed of John Kraft, Magda Colberg, Anne Gavin, and Lawrence Buck suggested changes and approved the final 30 items of each type to be experimentally tried out. Items were administered in the

following series on the dates specified.

Item Type

Word Selection
Analogies
Evaluation of Comparisons

Test 500 Series

21010
24010
24011

Date of Administration

May 1976
January 1977
January 1977

Results

Item analysis results from the three PACE administrations during which the experimental item types were tried out are summarized in Tables 1 and 2 and in Figures 1, 2 and 3. In all cases the statistics for the Comprehension item type used operationally to measure judgment can be compared with those for the experimental item types. It is important to note that the operational Judgment (Comprehension) sections from Series 24010 and 24011 are identical whereas this section in Series 21010 consists of a completely different set of Comprehension items. A different group of applicants is, of course, represented by each of the three series. Thus statistical variations among all four item types as well as between the operational and experimental item-types can be examined and compared relative to the variations which can occur when the same item type, or items themselves, are administered to different groups of people.

Table 1 shows that the statistics for the Comprehension items are quite stable; differences among the three series are not statistically significant. The test means, standard deviations, average p's and average r's for Word Selection and Evaluation of Comparisons were essentially the same as those for Comprehension. Analogies behaved somewhat differently. The mean for these items was about 1.5 points lower (18.51) than for the other item types, the standard deviation was almost 1 point larger (5.3), the average p was, of course lower (.62) and the average r of .40 was .06 higher. The internal consistency (KR_{20}) was also higher by .1 than those values for the other item types.

The Analogies items, however, had significantly higher correlations with the other constructs (especially the reasoning constructs of Induction and Deduction) than did Word Selection, Evaluation of Comparisons and Comprehension (see Table 2).

Figures 1, 2 and 3 present frequency distributions of the p and r values for the three experimental item types as well as for the Comprehension items in the three series. An inspection of these figures reveals that the item statistics covered the ranges of p and r quite well and the distributions for the experimental items were not overly far from the distributions for the Comprehension items. The distributions for Analogies are probably the most satisfying. There were at least three items in the p-value categories .2, .3 ... .8 and no items in the extremely easy .9 category. The r-values for this item type were also quite satisfactory; 18 of the 30 items had r's $>.4$ indicating that they successfully separated high and low scorers on the test.

Conclusions

The chief conclusion to be drawn from the foregoing research is that no clearly satisfactory replacement for the Comprehension item type has been found. Item statistics for Word Selection and Evaluation of Comparisons were (essentially) the same as those for Comprehension. One of the expectations for Word Selection and Evaluation of Comparisons was that their correlations with the Verbal section of PACE would be low. These correlations were indeed lower than the correlation between Verbal and Comprehension items. The differences in correlation (.03) were statistically significant because the sample sizes were larger ($\sim$2300) but such a reduction of a correlation in the .60's may not be of practical significance. In other words it may not be worthwhile to abandon the Comprehension item type in favor of either of these two new item types, because they, too, have a great deal of verbal variance. Correlations between the two experimental item types and the Comprehension item type were no higher than their correlation with Verbal items, indicating that Word Selection and Evaluation of Comparisons are no more predictive of judgment ability than they are of verbal ability. Lastly, the distribution of

TABLE I

Summary of Data for 3 Item Types and Comprehension Items - Test 500

Series 21010	$\bar{x}$	S	$\bar{p}$	$\bar{r}$	KR 20	N
Word Selection	20.06	4.09	.6687	.3264	.6870	1639
Comprehension	19.84	4.46	.6613	.3406	.7201	1639
Series 24011						
Evaluation of Comparisons	20.02	4.36	.6673	.3416	.7302	2276
Comprehension	19.98	4.44	.6660	.3310	.7234	2276
Series 24010						
Analogies	18.51	5.30	.6170	.4032	.8192	2299
Comprehension	20.08	4.44	.6693	.3346	.7262	2299

TABLE 2

Correlations - New Item Types and Comprehension with Other PACE Constructs - Test 500

Series 21010	Verbal	Compr.	Ind.	Ded.	Quant.
Word Selection	.6395	.6465	.4503	.5360	.3873
Comprehension	.6732	--	.4493	.5446	.4270
Series 24011					
Evaluation of Comparisons	.6797	.6704	.4531	.5672	.4139
Comprehension	.7065	--	.4298	.5696	.4636
Series 24010					
Analogies	.7275	.6403	.5335	.6381	.4688
Comprehension	.6995	--	.4048	.5842	.4484

p-value intervals	Series 21010	Series 24011	Series 24010
.20-.29	X	X	XXX O
.30-.39	XX O	XX OO	XXX O
.40-.49	XX OOOO	XXX OO	XXXXX O
.50-.59	XXXXXX OOOOOOO	XXXXX OOOOO	XXX OOOOOOO
.60-.69	XXXX OOOOO	XXXXX OOOOOOO	XXX OOOOOOO
.70-.79	XXXXXXX OOOOOOO	XXXXX OOOOOO	XXXXXX OOOOO
.80-.89	XXXX OOOO	XXXXXXX OOOOOOO	XXXXXXX OOOOOO
.90-.99	XXXX OO	XX O	OO
	X Word Selection O Comprehension	X Evaluation of Comparisons O Comprehension	X Analogies O Comprehension

Figure 1 Frequency distributions of p-values for three experimental item types and Comprehension items in three PACE series

r-value intervals	Series 21010	Series 24011	Series 24010
<.20	XXX OO	XXX OO	XX OO
.20-.29	XXXXXXXXX OOOOOO	XXXXXXXX OOOOOOOO	XXX OOOOOOOO
.30-.39	XXXXXXXXXXX OOOOOOOOOOOOOOO	XXXXXXXXX OOOOOOOOOOOOO	XXXXXXX OOOOOOOOOOOOOOO
.40-.49	XXXXXXX OOOOOOO	XXXXXXXXX OOOOOO	XXXXXXXXXXXXX OOOO
.50+		X O	XXXXX O
	X Word Selection O Comprehension	X Evaluation of Comparison O Comprehension	X Analogies O Comprehension

Figure 2 Frequency distributions of r-values for three experimental item types and Comprehension items in three PACE series.

p - and r - values for these two item types give further evidence that, relative to these particular 30 items, they may be equivalent, but not superior, to Comprehension.

Analogies as an item-type by itself presents item statistics which are a little more impressive, but these of course could be a function of the particular set of items included. Analogies had a higher correlation with the Verbal (.73), Induction (.53) and Deduction (.64) sections than any of the other three item types. Analogies, in their general conception, have frequently been associated factorially with Verbal Comprehension and Reasoning constructs. The specific form of analogy employed here (simple vocabulary but difficult relationships to match precisely) was intended to minimize verbal variance and maximize the task of evaluation or judgment in choosing the correct alternative. In spite of efforts to create such a situation, the analogy items, although correlated with Comprehension .64, are still more strongly associated with Verbal and Reasoning abilities than are the other item types for assessing judgment.

Empirical evidence for a Judgment factor, although present in the literature, is relatively weak and inconsistent. This, of course, made it difficult to find item types which could reasonably be expected to tap an ability called judgment. The three new item types proposed in this paper were chosen after carefully weighing statistics against feasibility of construction and other factors. A number of tests appeared to be promising in one study, only to be disqualified by the results of another. Other statistically appropriate measures would not lend themselves to a multiple choice format. An adequate range of item difficulty or variety of subject matter did not seem possible with still others. The three item types chosen - (1) Word Selection, (2) Evaluation of Comparisons and (3) Analogies seemed to have the best statistical evidence to associate them with judgment while minimizing other factor variance. They also exhibited the best possibilities for item variety, documentation of a defensible correct answer, appropriate subject matter for item content, and minimization of Verbal factor variance. Lastly, they best satisfied the requirements set forth in the PACE operational definition of judgment.

As previously stated, however, this preliminary tryout and analysis has not succeeded in presenting conclusively a good replacement for the Comprehension item type. Perhaps, as is believed by some authors (Northrop, 1976) judgment is an ability which is too complex and pervasive to be reliably measured in a pure factor sense.

REFERENCES

Adkins, D. C., & Lyerly, S. B. Factor analysis of reasoning tests. Chapel Hill, N. C.: University of North Carolina Press, 1952.

Guilford, J. P., Hertzka, A. F., Berger, R. M., & Christensen, P. R. A factor-analytic study of evaluative abilities: I. Hypotheses and description of tests. University of Southern California Psychological Laboratory Report 7, 1952.

Guilford, J. P., Hertzka, A. F., Christensen, P. R. A factor-analytic study of evaluative abilities: II. Administration of tests and analysis of results. University of Southern California Psychological Laboratory Report 9, 1953.

Guilford, J. P., & Hoepfner, R. The analysis of intelligence. New York: McGraw-Hill, 1971.

Nihira, K., Guilford, J. P., Hoepfner, R., & Merrifield, P. R. A factor analysis of the semantic evaluation abilities. University of Southern California Psychological Laboratory Report 32, 1964.

Northrop, L. C. The definition and measurement of judgment. Washington, D. C.: U. S. Civil Service Commission, 1976.

U. S. Civil Service Commission. Preparing for the Professional and Administrative Career Examination. Washington, D. C.: Author, 1976.